DOCUMENT RESUME

ED 423 303 TM 029 111

AUTHOR Huynh, Cam-Loi
TITLE Comparable Confidence Intervals for Multi-Sample and Replication Studies.
SPONS AGENCY Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE 1998-04-00
NOTE 48p.
CONTRACT 332-1665-34
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Estimation (Mathematics); Foreign Countries; *Research Methodology; *Sampling
IDENTIFIERS *Confidence Intervals (Statistics); *Research Replication

ABSTRACT

When the same parameters are estimated by data from several independent samples, it may happen that, for any pair of samples, even though the test for parameter discrepancy is statistically significant, the two individual confidence intervals overlap. To overcome this potential contradiction, a new type of one-sample confidence intervals is developed. Their evaluation will lead to the same statistical decisions reached by the two-sample test for parameter discrepancy. Moreover, the simultaneous decisions on parameter estimation, statistical inference, and directional prediction can be made with specified confidence coefficients and error rates by simply comparing a pair of comparable confidence intervals. In contrast with conventional confidence intervals, the comparable new confidence intervals have narrower widths, disjoint or overlap depending on whether the parameter discrepancy is statistically significant or not. The proposed procedure can be applied to both simple and multiple a-priori comparisons of means, proportions, and correlation coefficients. Due to its mathematical simplicity, the method should be valuable for research practitioners and quite suitable to be taught in courses of research methods in the behavioral and social sciences. An appendix explains the derivation of the formulas in Table 1. (Contains 3 tables and 49 references.) (Author/SLD)

Comparable Confidence Intervals for Multi-sample and Repication Studies

Cam-Loi Huynh

University of Manitoba

Running head: Comparable Confidence Intervals

Author Notes: This research was partially supported by a grant from the Social Sciences and Humanities Research Council of Canada (Act. No. 332-1665-34).
Correspondence concerning this article should be addressed to Cam-Loi Huynh, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2 or by fax at (204) 474-7599. Electric mail may be sent via Internet to Huynh@cc.umanitoba.ca.

Comparable Confidence Intervals for Multi-sample and Repication Studies

Abstract

When the same parameters are estimated by data from several independent samples, it may happen that, for any pair of samples, even though the test for parameter discrepancy is statistically significant, the two individual confidence intervals overlap. To overcome this potential contradiction, a new type of one-sample confidence intervals is developed. Their evaluation will lead to the same statistical decisions reached by the two-sample test for parameter discrepancy. Moreover, the simultaneous decisions on parameter estimation, statistical inference and directional prediction can be made with specified confidence coefficients and error rates by simply comparing a pair of comparable confidence intervals. Contrasting to the corresponding conventional confidence intervals, the comparable confidence intervals have narrower widths, disjoint or overlap depending on whether the parameter discrepancy is statistically significant or not. The proposed procedure can be applied to both simple and multiple a-priori comparisons of means, proportions and correlation coefficients. Due to its mathematical simplicity, the method should be valuable for research practitioners and quite suitable to be taught in courses of research methods in the behavioral and social sciences.

Comparable Confidence Intervals for Multi-sample and Repication Studies

There are several circumstances in which the comparison of individual confidence intervals (for ϕ_i, $\underline{i}$ = = 1, 2, ..., $\underline{k}$) across a series of independent samples is needed. The comparison may be conducted jointly with the evaluation of the associated significance tests ($\underline{H_0}$: $\phi_i = \phi_0$, $\underline{i}$ = 1, 2, ..., $\underline{k}$) or confidence intervals of the parameter discrepancy (δ, where $\delta = \phi_i - \phi_{i'}$, $\underline{i} \neq \underline{i}'$ = 1, 2, ..., $\underline{k}$). First, the comparison of confidence intervals of the true parameter, associated with the same hypotheses but obtained under various research conditions (e.g., with different sample sizes and sample variances in multi-sample studies) will help identifying not only the statistically significant results but possibly also the practically, or clinically, important hypothetical conjectures. Usually such comparisons can be performed by means of simple bar or line graphs. For example, the graph contains confidence intervals drawn horizontally one on top of another and a vertical line representing the hypothetical value of the parameter. The confidence interval that is away furthest from the vertical line in the predicted direction may be chosen to indicate the range of both statistically and clinically significant effect (Borenstein, 1994). Secondly, sometimes confidence intervals may be more informative than statistical tests in the evaluation and comparison of the statistical results. For example, a student's performance in a national standards test of Mathematics is deemed unsatisfactory. One would reach such a conclusion more convincingly if it can be shown that the two confidence intervals for the individual and national true means are separable, namely, even the upper bound of the former falls below the lower bound of the latter. Clearly, for this type of single-subject analysis, visual methods such as the comparison

of confidence intervals are as necessary as, if not more meaningful than, the p-value or an index of the statistical power of the test. Thirdly, if the research objective is to estimate the typical range of the parameter of interest on the basis of multi-sample studies then confidence intervals of the parameter discrepancy may not be relevant. One comparing individual intervals for the parameter itself will identify the extent of sampling fluctuations and, as a result, obtain a more precise estimate of the true parameter range. This is especially fitting if estimates of the parameter discrepancy are found statistically significant. Hsu (1994) gives the following example, "two data sets may give rise to the two confidence intervals $\mu_i \in 1 \pm 0.2$ months and $\mu_i \in 10 \pm 2$ months, which convey very different information about μ, yet the same p-value associated with H_{i0}" (p. 4), where H_{i0}: $\mu_i = 0$. Note that the test of means difference and the two individual mean tests may be all statistically significant (possibly at approximately the same p values), but the two individual confidence intervals have very different widths. Depending on the units of measurement and what the parameter (μ) represents, the researcher would prefer one but not the other confidence interval for the estimate of the parameter range. Considerations as such are often overlooked if one computes only the test of means difference[1]. Moreover, the computation of the significance test is not necessary if one wants to estimate the p-value of the test statistic and the power of the test. It will be shown that, given the information of a confidence interval, it is possible to recover the p-value of the associated test statistic, and the power of the test is the same as the power of the confidence interval. Last but not least, the confidence interval can be used to identify the directions of the parameter and the parameter discrepancy. For the test of H_{i0}: $\phi_i = 0$, if the confidence interval for ϕ_i is entirely to the left of 0 then $\phi_i < 0$ whereas if it is completely to the right of 0 then $\phi_i > 0$. Following the procedures of two-tailed directional tests (Kaiser,

1960; Shaffer, 1972; and Leventhal and Huynh, 1996a , 1996b), one can compute the risk of making directional decisions (Type III error rate).

But how can confidence intervals computed for different samples be compared? It has been recognized that, within each sample, the same statistical decision can be reached by either conducting a statistical test (e.g., an evaluation of the p-value of the test statistic) or by examining if the hypothetical value of the parameter falls within the corresponding confidence limits. The match between statistical tests and confidence intervals for statistical inference is desirable and plays a major role for advocating the use of confidence intervals (Natrella, 1960). However, the comparison of the individual confidence intervals obtained for the single parameters, say means, may or may not reproduce the same statistical decisions on the statistical significance of means difference across different samples. This is because, for any pair of means, the tests for individual means and means discrepancy are based on different estimates of the standard errors, and on t statistics with different degrees of freedom if the population variances are unknown.

There is a growing interest in confidence intervals among applied researchers[2]. It is expected that confidence intervals will be the standard method for statistical inference in social and behavioral sciences. Although statistical decisions based on post-hoc multiple comparisons of confidence intervals for the means have been discussed, there is still a need for a systematic study on the procedures and conditions for comparing confidence intervals associated with pre-planned tests of means, proportions and correlations[3]. But the methods for evaluating confidence intervals are lacking. It is the purpose of this paper to fulfill this need. Moreover, it will be argued that the statistical decisions based on comparable confidence intervals will satisfy the golden rule of equivalent outcomes between hypothesis testing and

confidence interval evaluation, a hurdle that prevents the comparison of the conventional confidence intervals unless very stringent conditions are met (e.g., same sample sizes and population variances). The methods to compute and evaluate comparable confidence intervals will be discussed in the first three sections. Readers unconcerned with the methodological development but want a quick overview of the utility of the proposed procedures may be benefitted by reading the examples at the end of these sections before moving to the bulk of the article. Some related issues in the application of the proposed procedure and general conclusions will be drawn in the final section.

Background of the Study

The Problem

In the following discussion, all tests are two-tailed, evaluated at α, or the nominal significance level, so that $\alpha/2$ is the size of each tailed critical region. Consider $\underline{k}$ independent populations characterized by the parameters ϕ_i, $\underline{i} = 1, ..., \underline{k}$, for $\underline{k} \geq 2$, where ϕ may denote the mean (μ), proportion (π) or correlation coefficient (ρ). The development of the proposed method is based on the following "equivalency principle":

> If H_0: $\phi_i = \phi_{i'}$ is rejected at α in favor of H_A: $\phi_i \neq \phi_{i'}$ for any $\underline{i} \neq \underline{i}' = 1, ..., \underline{k}$, then the individual $100(1 - \alpha)\%$ two-tailed confidence intervals for H_0: $\phi_i = \phi_0$ and H_0: $\phi_{i'} = \phi_0$, where ϕ_0 be a hypothetical value of ϕ, should be separable, or nonoverlapping. On the other hand, if H_0: $\phi_i = \phi_{i'}$ is conceded then the individual confidence intervals for ϕ_i, $\underline{i} \neq \underline{i}' = 1, ..., \underline{k}$ conducted at the same level of α, overlap. Two confidence intervals are said to overlap if the upper (lower) bound of the confidence interval for the smaller (larger) parameter

estimate is located inside the confidence interval for the larger (smaller) parameter estimate. Otherwise, the confidence intervals are considered separable.

For simplicity, the following discussion is based on the tests of means in two-sample studies under the assumption of variance homogeneity ($\sigma^2_i = \sigma^2$, $\underline{i} = 1, 2$). Let $\underline{M_i}$ be the sample mean of the ith independent population with the corresponding sample size $\underline{n_i}$, $\underline{i} = 1, 2$, not necessarily equal. The $100(1 - \alpha)\%$ two-tailed individual confidence intervals for testing $H_{0,i}$: $\mu_i = \mu_0$ ($\underline{i} = 1, 2$) are computed by the conventional method as

$$(1) \qquad \underline{CI_i}: \underline{M_i} \pm \underline{Z_{1-\alpha/2}SE_i}, \ \underline{i} = 1, 2,$$

(two one-sample confidence intervals), where $\underline{SE_i} = \sigma/\sqrt{\underline{n_i}}$, $\underline{i} = 1, 2$ and $\underline{Z_{1-\alpha/2}}$ = the $(1 - \alpha/2)$th quantile of the standard normal distribution (such that $\underline{Z_{\alpha/2}} = -\underline{Z_{1-\alpha/2}}$). Without loss of generality, assuming $\underline{M_1} > \underline{M_2}$. The corresponding confidence interval associated with the test of $\underline{H_{0,d}}$: $\mu_1 - \mu_2 = 0$ is

$$(2) \qquad \underline{CI_d}: (\underline{M_1} - \underline{M_2}) \pm \underline{Z_{1-\alpha/2}SE_d},$$

(a two-sample confidence interval), where $\underline{SE_d} = \sigma\sqrt{\{(1/\underline{n_1}) + (1/\underline{n_2})\}}$. The subscript $\underline{d}$ represents the fact that the means difference is being assessed. If the confidence interval $\underline{CI_d}$ does not contain the value of zero then the null hypothesis $\underline{H_{0,d}}$ is rejected at the predetermined significance level α for two-tailed test. Otherwise, if $\underline{CI_d}$ does not contain zero, one fails to reject the null hypothesis $\underline{H_{0,d}}$.

It is possible that the two within-sample confidence intervals overlap even though the between-sample test for the parameter discrepancy is statistically significant at α. The

question of interest is how the individual confidence intervals in Equation 1 be compared so that the same statistical decision obtained for $\underline{H}_{0,d}$ can be reached according to the "equivalency principle" stated above.

A Solution

Let $\underline{CI}_{i,U}$, and $\underline{CI}_{i,L}$, $\underline{i}$ = 1, 2, denote the upper and lower bounds of the individual confidence intervals, respectively, and ε_i, $\underline{i}$ = 1, 2, be any small, positive constant. Since $\underline{M}_1 > \underline{M}_2$, $\underline{CI}_{1,U}$ lies to the right of $\underline{CI}_{2,U}$. If $\underline{H}_{0,d}$ is rejected at $\alpha/2$ then the lower bound of $\underline{CI}_1$ is at least equal to the smaller mean ($\underline{M}_2$) so that $\underline{M}_1 - \underline{Z}_{1-\alpha/2}\sigma/\sqrt{\underline{n}_1} = \underline{M}_2 + \varepsilon_1$, or

$$(\underline{M}_1 - \underline{M}_2) - \underline{Z}_{1-\alpha/2}\sigma/\sqrt{\underline{n}_1} = \varepsilon_1, \tag{3}$$

and the upper bound of $\underline{CI}_2$ is at most equal to the larger mean ($\underline{M}_1$) so that $\underline{M}_2 + \underline{Z}_{1-\alpha/2}\sigma/\sqrt{\underline{n}_2} = \underline{M}_1 - \varepsilon_2$, or

$$(\underline{M}_1 - \underline{M}_2) - \underline{Z}_{1-\alpha/2}\sigma/\sqrt{\underline{n}_2} = \varepsilon_2. \tag{4}$$

Conditions (3) and (4) imply that when the test of means difference is statistically significant, the lower bound of the 100(1 - α)% confidence interval computed from the individual samples should be

$$\underline{CI}_{c,L}: (\underline{M}_1 - \underline{M}_2) - \underline{Z}_{1-\alpha/2}\sigma\{(1/\sqrt{\underline{n}_1}) + (1/\sqrt{\underline{n}_2})\}.$$

The subscript $\underline{c}$ stands for the correction on the conventional one-sample confidence intervals. By symmetry, it is easy to write the equation for its upper bound. In other words, if the difference of the two individual confidence intervals in Equation 1 is computed, the resulting 100(1 - α)% two-tailed confidence interval for the means difference would be specified as

(5) $$\underline{CI}_c: (\underline{M}_1 - \underline{M}_2) \pm \underline{Z}_{1-\alpha/2}\underline{SE}_c,$$

where $\underline{SE}_c = \underline{SE}_1 + \underline{SE}_2 = \sigma\{(1/\sqrt{\underline{n}_1}) + (1/\sqrt{\underline{n}_2})\}$. The width of the confidence interval in Equation 5 is larger than that in Equation 2 since $\{(1/\sqrt{\underline{n}_1}) + (1/\sqrt{\underline{n}_2})\} > \sqrt{\{(1/\underline{n}_1) + (1/\underline{n}_2)\}}$ for any positive constant $\underline{n}_i$, $\underline{i} = 1, 2$. This explains why the "equivalence principle" can be violated. To prevent this possibility, a modification of $\underline{SE}_i$, $\underline{i} = 1, 2$, that renders the equality $\underline{SE}_c = \underline{SE}_d$ must be found. As a solution, the critical values in Equation 1 is set to be $\underline{Z} = \underline{Z}_{1-\alpha/2}\underline{c}_f$, where $\underline{c}_f$ is the correction factor of the form

(6) $$\underline{c}_f = \frac{\underline{SE}_d}{\underline{SE}_c} = \frac{\sqrt{\{\underline{n}_1 + \underline{n}_2\}}}{\sqrt{\underline{n}_1} + \sqrt{\underline{n}_2}},$$

implying $\underline{c}_f < 1$. Upon replacing $\underline{Z}_{1-\alpha/2}$ by $\underline{Z}_{1-\alpha/2}\underline{c}_f$ in Equation 1, the conventional confidence intervals are revised to yield the following comparable confidence intervals

(7) $$\underline{CI}^*_i: \underline{M}_i \pm \underline{Z}_{1-\alpha/2}\underline{SE}^*_i,$$

where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, for $\underline{i} = 1, 2$. Clearly, the width of $\underline{CI}^*_i$ is narrower than that of $\underline{CI}_i$, for $\underline{i} = 1, 2$, respectively. To recapitulate, if the means difference is statistically significant at $\alpha/2$ then the comparable confidence intervals in Equation 7 are separable. This can be accomplished even if the two $100(1-\alpha)\%$ two-tailed confidence intervals in Equation 1 overlap.

An Alternative Solution

For the purpose of preserving the "equivalency principle", instead of adjusting the standard error of the estimates in computing $\underline{SE}^*_i$, $\underline{i} = 1, 2$, as above, one can modify the nominal significance level and revise the confidence coefficient of the conventional intervals

accordingly. The formula for computing a Type I probability corresponding to a 100(1 - α)% conventional confidence interval is given by

(8) $$\alpha = 2[1 - \Phi(Z_{1 - \alpha/2})],$$

where $\Phi(x) = Pr(X \leq x)$ = the probability of obtaining a standard normal value x = the area under the standard normal curve to the left of the point x. Therefore, the adjusted or comparative Type I error (α'), corresponding to the nominal level (α), can be derived as

(9) $$\alpha' = 2[1 - \Phi(c_f Z_{1 - \alpha/2})].$$

This enables the computation of comparative confidence intervals, defined as the 100(1 - α')% two-tailed conventional confidence interval of the form

(10) $$CI'_i: M_i \pm Z_{1 - \alpha'/2} SE_i.$$

Since $\Phi(Z_{1 - \alpha/2}) > \Phi(c_f Z_{1 - \alpha/2})$, α is smaller than α', implying that $Z_{1 - \alpha'/2} \leq Z_{1 - \alpha/2}$, the length of the 100(1 - α')% comparative confidence interval (CI'_i) is narrower than that of the 100(1 - α)% conventional confidence interval (CI_i) for i = 1, 2, respectively.

Example 1

Consider two independent samples drawn from a population having a known standard deviation of σ = 10, the first sample with n_1 = 36, M_1 = 25.5 and the second sample with n_2 = 25 and M_2 = 20. The 95% two-tailed individual confidence intervals for testing $H_{0,i}$: $\mu_i = \mu_0$ (i = 1, 2), for any value of μ_0, are given by the standard method as $CI_1 = 25.5 \pm (1.96)(10/6)$ = (22.23, 28.77) and $CI_2 = 20 \pm (1.96)(10/5)$ = (16.08, 23.92). These two conventional

confidence intervals overlap from 22.23 to 23.92, or about 26% of the confidence interval for μ_1 and 22% of the confidence interval for μ_2. However, the test of means difference ($\underline{H}_{0,d}$: $\mu_1 = \mu_2$) is statistically significant at $\alpha = .05$ (two-tailed) since its 95% two-tailed confidence interval is equal to $\underline{CI}_d = (25.5 - 20) \pm (1.96)(10)\sqrt{\{(1/36) + (1/25)\}} = (0.40, 10.60)$. The required correction factor is computed as $\underline{c}_f = \sqrt{\{36 + 25\}}/(\sqrt{36} + \sqrt{25}) = 0.71$. Hence, the corresponding comparable confidence intervals are $\underline{CI}_1 = 25.5 \pm (1.96)(0.71)(10/6) = (23.18, 27.82)$ and $\underline{CI}_2 = 20 \pm (1.96)(.71)(10/5) = (17.22, 22.78)$. As expected, the 95% two-tailed comparable confidence intervals are separable. The comparative Type I error is $\alpha' = 2[1 - \Phi((.71)(1.96))] = 2[1 - 0.91798] = 0.164$ and the comparative confidence intervals are $\underline{CI}'_1 = (24.34, 26.66)$ and $\underline{CI}'_2 = (18.61, 21.40)$. Therefore, the following three statistical decisions can be made for the given data: (i) the conventional 83.6% two-tailed confidence intervals for testing $\underline{H}_{0,i}$: $\mu_i = \mu_0$ (i = 1, 2) are disjointed, (ii) the 95% two-tailed comparable confidence intervals are separable, and (iii) the test of means difference ($\underline{H}_{0,d}$: $\mu_1 = \mu_2$) is statistically significant at $\alpha = .05$ (two-tailed)[4].

Comments

Although the evaluation of $\underline{CI}^*_i$ and $\underline{CI}'_i$ would yield outcomes that satisfy the "equivalency principle" in statistical decisions, the two types of statistical intervals are not identical. It is recommended that the comparable confidence intervals be used at the expense of the comparative confidence intervals for a-prior pairwise comparisons. First of all, because $\underline{CI}'_i$'s are computed on the basis of both the correction factor $\underline{c}_f$ and comparative significance level α' as shown in Equation 9, they may be overadjusted for the statistical significance of the between-sample test. Moreover, since α must be set in advance, it is more natural to compute the comparable confidence intervals with the $(1 - \alpha)$ coefficient than the comparative

confidence intervals with a corrected coefficient $(1 - \alpha')$.

The significance of the test statistic is measured by its p-value. Since the p-value is the Type I error probability computed on the basis of sample data, its formula is the same as Equation 8 upon replacing $Z_{1 - \alpha/2}$ by the test statistic itself (Welsh, 1996). Let α_p denote the p-value for the test of means difference. It can be computed as

$$\alpha_p = 2[1 - \Phi\{(M_1 - M_2)/SE_d\}],$$

where $SE_d = \sigma\sqrt{\{(1/n_1) + (1/n_2)\}}$. The $100(1 - \alpha_p)\%$ conventional confidence interval contains the value of 0 as its lower bound for $M_1 > M_2$ (or its upper bound if $M_1 < M_2$). It may be called the significant confidence interval for estimating the parameter discrepancy and is computed as

$$CI_s\text{: } (M_1 - M_2) \pm Z_{1 - \alpha p/2}SE_d = (0, 2(M_1 - M_2)),$$

since $Z_{1 - \alpha p/2} = |M_1 - M_2|$ for $M_1 > M_2$. Although the p-value of the test statistic is generally meaningful, for the simultaneous evaluation of H_{0i} and H_d, the significant confidence interval may not be relevant and will not be discussed further in this paper[4].

Using Comparable Confidence Intervals for Parameter Estimation and Hypothesis Testing in Two-sample Studies

In the following, a procedure for two-sample tests of means, proportions and correlation coefficients will be discussed in the context of a simultaneous assessment of within-sample and between-sample tests of the same population parameter. The discussion will be carried out for three popular types of two-tailed tests: (i) nondirectional tests evaluated at $\alpha/2$, (ii) nondirectional tests with unequal allocation of Type I error rates, and (iii) nondirectional within-sample tests and a directional between-sample test, all are evaluated at $\alpha/2$.

Nondirectional Two-Tailed Tests with Symmetric Critical Regions

Procedure

Suppose the researcher wants to estimate the range of each ϕ_i, $i = 1, 2$, and simultaneously, to test the difference between ϕ_1 and ϕ_2. For these purposes, in the following procedure, it suffices to evaluate only the one-sample comparable confidence intervals. The testing procedure consists of the following steps:

(a) Specifying the hypotheses:

(Within-sample tests): $H_{0,i}$: $\phi_i = \phi_0$ vs. H_A: $\phi_i \neq \phi_0$, $i = 1, 2$; where ϕ_0 can be any value of interest.

(Between-sample test): $H_{0,d}$: $\phi_1 - \phi_2 = 0$ vs. $H_{A,d}$: $\phi_1 - \phi_2 \neq 0$.

(b) Computing the $100(1 - \alpha)\%$ two-tailed comparable confidence intervals (CI^*_i):

$$CI^*_i: \hat{\phi}_i \pm CV_{1-\alpha/2} SE^*_i, \; i = 1, 2,$$

where $SE^*_i = c_r SE_i$, CV is the critical value of the test statistic and $\hat{\phi}_i$ represents the sample estimate of ϕ_i.

(c) Making two statistical decisions at the same significance level of $\alpha/2$:

(i) Reject $H_{0,i}$ if the comparable confidence interval CI^*_i does not contain ϕ_0 and decide that ϕ_i is probably not equal to ϕ_0. Otherwise, concede $H_{0,i}$, $i = 1, 2$ and assume that ϕ_i is equal to ϕ_0.

(ii) Reject $H_{0,d}$ if the comparable confidence intervals CI^*_1 and CI^*_2 are disjointed and decide that the difference in the two estimates of ϕ_1 and ϕ_2 are statistically significant. Otherwise, concede $H_{0,i}$, $i = 1, 2$, and assume that the difference in the two parameter

estimates are not statistically significant.

For the two-sample tests of means, proportions and correlation coefficients, the general form of the correction is found to be

$$(11) \qquad \underline{c}_f = \frac{\sqrt{\{\underline{a} + \underline{b}\}}}{\sqrt{\underline{a}} + \sqrt{\underline{b}}},$$

where $\underline{a}$ and $\underline{b}$ are functions of variances and/or the number of cases. The results are summarized in Table 1 and their derivations are given in the Appendix.

Insert Table 1 about here

Equation 11 can be simplified to be $\underline{c}_f = \sqrt{\{1 + \underline{r}\}}/(1 + \sqrt{\underline{r}})$ where $\underline{r} = \underline{a}/\underline{b}$, the values of $\underline{a}$ and $\underline{b}$ can be defined such that $\underline{a} > \underline{b}$ and $\underline{r} > 1$. Since $\underline{r}$ represents the ratio of sample variances and/or cases, its values would likely be from 1 to 10 in most practical research situations. In these circumstances, the range of $\underline{c}_f$ would be from .71 (when $\underline{r} = 1$) to .80 (when $\underline{r} = 10$). Table 2 provides the comparative confidence coefficients $(1 - \alpha')$ for the conventional confidence intervals corresponding to the comparable confidence intervals computed with the nominal α levels for $\underline{Z}$ and $\underline{t}$ tests under the assumption of variance homogeneity. For example, if $\underline{H}_{0,d}$ is statistically significant then, given a correction factor of .70 from a $\underline{Z}$ distribution, both the 95% comparable confidence interval and the 83% conventional confidence interval have separable limits. Given a correction factor of .76, if $\underline{H}_{0,d}$ is statistically significant then an 80% conventional (or comparative) confidence interval and the 90% comparable confidence intervals for the $\underline{t}$ test with 10 degrees of freedom have disjointed confidence bounds.

Insert Table 2 about here

Example 2

Two samples are drawn randomly from a population having unknown variance (with $n_1 = 31$, $M_1 = 28$ and $S^2_1 = 144$; and $n_2 = 16$, $M_2 = 20$ and $S^2_2 = 62$, respectively). Suppose the test of variance homogeneity (O'Brien, 1981) is statistically significant at $\alpha = .05$ (This is the Case 1d in the Appendix). The 95% two-tailed individual confidence intervals for testing $H_{0,i}$: $\mu_i = \mu_0$ ($i = 1, 2$), for any value of μ_0, are given by the standard method as $CI_1 = (23.60, 32.40)$ and $CI_2 = (15.80, 24.20)$. The overlap from 23.60 to 24.20 represents nearly 7% of the interval length of either CI_1 or CI_2. However, the test of means difference ($H_{0,d}$: $\mu_1 = \mu_2$) is statistically significant at $\alpha = .05$ (two-tailed) according to the 95% two-tailed confidence interval for the mean discrepancy (δ) $CI_d = (2.11, 13.89)$. In computing CI_d, for $f^* = 42.20$ (the degree of freedom according to Satterwaite-Welch approximation for t tests), the critical value of $t_{f^*, 1 - \alpha/2}$ is found to be 2.02 (say, by using the command TINV(.025, 42.20) in the SAS computer program, SAS Institute Inc., 1990). The required correction factor is $c_f = 0.71$ and the adjusted within-sample standard errors are $SE^*_1 = 1.53$ and $SE^*_2 = 1.39$. Hence, the corresponding comparable confidence intervals are $CI^*_1 = (24.88, 31.12)$ and $CI^*_2 = (17.03, 22.97)$. As expected, the 95% two-tailed comparable confidence intervals are separable. The comparative Type I error rate is $\alpha' = 2[1 - T_{f^*}((.71)(2.018))] = 2[1 - 0.920] = 0.16$ and the associated comparative confidence intervals are also disjointed, being $CI'_1 = (25.83, 30.17)$ and $CI'_2 = (17.98, 22.02)$.

Nondirectional Two-Tailed Tests with Asymmetric Critical Regions

Procedure

Suppose the researcher wants to test the difference between ϕ_1 and ϕ_2 and to estimate the range of ϕ_i at different confidence coefficients, say $(1 - \alpha_1)$ and $(1 - \alpha_2)$. The relevant hypotheses are specified below.

(Within-sample tests):

$$\underline{H}_{0,1}: \phi_i = \phi_0 \text{ vs. } \underline{H}_A: \phi_1 \neq \phi_0, \text{ evaluated at } \alpha_1/2,$$
$$\underline{H}_{0,2}: \phi_i = \phi_0 \text{ vs. } \underline{H}_A: \phi_1 \neq \phi_0, \text{ evaluated at } \alpha_2/2,$$

(Between-sample test): $\underline{H}_{0,d}: \phi_1 - \phi_2 = 0$ vs. $\underline{H}_{A,d}: \phi_1 - \phi_2 \neq 0$, evaluated at $\Gamma/2$.

In this case, the significance levels of α_1 and α_2 are predetermined but Γ is a function of α_1 and α_2. We now show how a value of Γ can be determined. Following the same argument leading to Equation 5, the resulting confidence interval for the parameter difference is of the form

(12) $$\underline{CI}_c: (\hat{\phi}_1 - \hat{\phi}_2) \pm \underline{Z}_{1-\Gamma/2}\underline{SE}_c = (\hat{\phi}_1 - \hat{\phi}_2) \pm \underline{Q},$$

where $\underline{Q} = [\underline{Z}_{1-\alpha 1/2}\sqrt{\underline{n}_2} + \underline{Z}_{1-\alpha 2/2}\sqrt{\underline{n}_1}]/\sqrt{\underline{N}}$, $\underline{N} = \underline{n}_1 + \underline{n}_2$. Hence, the required Type I error rate for the between-sample test becomes

(13) $$\Gamma = 2[1 - \Phi(\underline{Q})]$$

where the function $\Phi(x)$ has been defined previously. From equation 13, $\underline{Q}$ is the inverse

function of the standard normal distribution evaluated at $(1 - \Gamma/2)$, i..e, $Q = \Phi^{-1}(1 - \Gamma/2) = \underline{Z}_{1 - \Gamma/2}$. The corresponding comparable confidence intervals are given by

$$\underline{CI}^*_i: \hat{\phi}_i \pm \underline{Z}_{1 - \Gamma/2}\underline{SE}^*_i = \hat{\phi}_i \pm \underline{Q}\underline{SE}^*_i, \tag{14}$$

where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, for $\underline{i} = 1, 2$, are the same as derived for symmetric confidence intervals.

Alternatively, for these sample sizes, by declaring significance whenever a $100(1 - \alpha_1)\%$ conventional confidence interval for μ_1 does not overlap a $100(1 - \alpha_2)\%$ confidence interval for μ_2, one decides that the two-sample test of $\mu_1 - \mu_2$ is statistically significant at $\Gamma/2$. Otherwise, by declaring these two one-sample confidence intervals overlap, one concedes the null hypothesis of $\mu_1 - \mu_2$ at $\Gamma/2$. The comparative Type I error rate and the related comparative confidence intervals are

$$\alpha' = 2[1 - \Phi(\underline{c}_f\underline{Q})],$$

and,

$$\underline{CI}'_i: \underline{M}_i \pm \underline{Z}_{1 - \alpha'/2}\underline{SE}_i,$$

respectively.

Example 3

For the data in Example 1, the correction factor is computed to be $\underline{c}_f = .071$. Suppose $\alpha_1 = .05$ and $\alpha_2 = .10$ then $\underline{Q} = [(1.96)\sqrt{25} + (1.645)\sqrt{36}]/\sqrt{\{36 + 25\}} = 2.52$ and $\Gamma = 2[1 - \Phi(2.52)] = 2(1 - .994) = .012$. The 98.8% two-tailed confidence interval for the difference scores is $\underline{CI}_d$: $(25.5 - 20) \pm (2.52)(10)\sqrt{\{(1/36) + (1/25)\}} = (-1.06, 12.06)$, implying that the means difference is not statistically significant. As expected, the 95% two-tailed comparable confidence intervals for the two samples, namely $\underline{CI}^*_1$: $25.5 \pm (2.52)(0.71)(10/6) = (22.52,$

28.48) and $\underline{CI}^*_2$: $20 \pm (2.52)(0.71)(10/5) = (16.42, 23.58)$, respectively, are not separable.

Comments

The symmetric tests in the previous section can be considered as a special case of the asymmetric tests in which $\alpha = \alpha_1 = \alpha_2$. For the asymmetric case, the two within-sample tests would have unequal statistical powers with respect to the alternative hypotheses that are equivalent in magnitude. This might be desirable if the likelihood to detect the hypothetical value (ϕ_0) in one sample is greater than in the other.

Directional Comparable Confidence Intervals

To this point, the individual comparable confidence intervals are used to assess the within-sample and between-sample tests simultaneously. These testing procedures are nondirectional since they do not provide the answers to, say, the following questions: "Which of ϕ_i ($\underline{i} = 1, 2, ..., \underline{k}$) is better? What would be the risk in making such a selection?". A procedure developed by Kaiser (1960) and modified by Shaffer (1972) and Leventhal and Huynh (1996) can be used to address this problem. With the introduction of the directional hypotheses for the two-tailed test of means difference, the one-sample comparable confidence intervals can be used yet for another purpose, namely, to predict the direction of the parameter difference. An important concept in evaluating directional hypotheses is the Type III error rate (γ) or "the risk of getting the direction wrong upon the rejection of the null hypothesis". It constitutes a component in the formula for the statistical power,

$$\text{Corrected Power: } \pi(\mu)^* = 1 - \beta - \gamma,$$

where β and γ are the probabilities of making a Type II error and Type III error, respectively. Computationally, the Type III error rate is represented by the tailed area opposite to the predicted direction under the distribution of the alternative hypothesis. Consider the two-sample test of means discrepancy ($\delta = \mu_1 - \mu_2$). Without loss of generality, suppose $\delta > 0$. Let $\underline{Z}_2 = \underline{Z}_{1-\alpha/2} - \delta_A/\underline{SE}$ and $\underline{Z}_1 = -(\underline{Z}_1 + \delta_A/\underline{SE})$, where δ_A is the assumed value of the mean difference under the alternative hypothesis; $\underline{Z}_2$ and $\underline{Z}_1$ represent the right and left limits of the central range (or region of accepting the null hypothesis) but computed under the alternative hypothesis, respectively. For the nondirectional two-tailed test using the $\underline{Z}$ statistic, the conventional power is defined as

$$\text{Conventional Power: } \pi(\mu) = 1 - \beta = 1 + \Phi(\underline{Z}_1) - \Phi(\underline{Z}_2),$$

(DeGroot, 1975, pp. 404-405; Zehna, 1970, p. 447). Hence, $\beta = \Phi(Z_2) - \Phi(\underline{Z}_1)$ and $\gamma = \Phi(Z_1)$ so that

$$\pi(\mu)^* = \pi(\mu) - \gamma = 1 - \Phi(\underline{Z}_2).$$

One may wish to know that the Type III error is always less than $\alpha/2$ (Kaiser, 1960, p. 164; Leventhal and Huynh, 1996a, p.284)[5].

Procedure

(a) Specifying the hypotheses:

The within-sample tests remain the same but the between-sample tests are based on three sets of hypotheses:

(Within-sample tests): $\underline{H}_{0,i}$: $\phi_i = \phi_0$ vs. $\underline{H}_A$: $\phi_i \neq \phi_0$, $\underline{i} = 1, 2$; where ϕ_0 can be any value of interest.

(Between-sample test): $\underline{H}_{d1}$: $\phi_1 - \phi_2 < 0$, $\underline{H}_{d2}$: $\phi_1 - \phi_2 = 0$ (null hypothesis), and $\underline{H}_{d3}$:

$\phi_1 - \phi_2 > 0$

(c) Computing the 100(1 - α)% two-tailed comparable confidence interval ($\underline{CI}^*_i$) (same as under the nondirectional procedure with $\alpha/2$)

(b) Making two statistical decisions at the significance level of $\alpha/2$. The first decision (i) remains unchanged. The second decision is modified as follows:

(ii) If the comparable confidence intervals $\underline{CI}^*_1$ and $\underline{CI}^*_2$ are disjointed, reject $\underline{H}_{d2}$ in favor of $\underline{H}_{d1}$ if $\hat{\phi}_1 < \hat{\phi}_2$ (or reject $\underline{H}_{d2}$ in favor of $\underline{H}_{d3}$ if $\hat{\phi}_1 > \hat{\phi}_2$, where $\hat{\phi}_i$ is the sample estimate of ϕ_i, $\underline{i}$ = 1, 2) with a Type III error rate (γ) less than $\alpha/2$. Otherwise, if $\underline{CI}^*_1$ and $\underline{CI}^*_2$ overlap, concede $\underline{H}_{0,i}$, $\underline{i}$ = 1, 2, and assume that the difference in the two parameter estimates are not statistically significant.

Suppose the null hypothesis ($\underline{H}_{d2}$) of the between-sample test is evaluated at α, where α = the probability that at least one of the alternative will be rejected, given that the null hypothesi is true (called "the overall significance level" by Shaffer, 1972, p. 196; and Leventhal and Huynh, 1996a, p.279). Then, the one-tailed tests $\underline{H}_{d1}$ and $\underline{H}_{d2}$ should be conducted at $\alpha/2$ (or their 100(1 - α)% conventional confidence intervals be evaluated). Since only one of the one-tailed test can be statistically significant at $\alpha/2$, if any, the researcher can decide whether ϕ_1 is less than, more than, or equal to ϕ_2 according to whether $\underline{H}_{d1}$, or $\underline{H}_{d2}$, or neither, is statistically significant at $\alpha/2$. As stated above, the probability of making a mistake in deciding the direction is called Type III error ($\leq \alpha/2$). Besides these changes, the computation of the one-sample conventional confidence intervals, correction factor, comparable confidence intervals, comparative error rates and comparative confidence intervals follow the same formulas derived for the nondirectional tests.

Example 4

Recall that in Example 1 the test of means difference ($\underline{H}_{d2}$: $\phi_1 - \phi_2 = 0$) is significant at $\alpha_t = .05$ (two-tailed). Had the hypotheses $\underline{H}_{d1}$: $\phi_1 - \phi_2 < 0$ and $\underline{H}_{d3}$: $\phi_1 - \phi_2 > 0$ been evaluated, the resulting 97.5% one-tailed confidence intervals are

$$\underline{CI}_{d1}: (\mu_1 - \mu_2) \in (-\infty, 10.60), \text{ and } \underline{CI}_{d3}:(\mu_1 - \mu_2) \in (0.40, \infty)$$

Since the confidence interval of $\underline{CI}_{d3}$ does not contain zero, one decides at $\alpha/2$ that $\underline{H}_2$ is rejected in favor of $\underline{H}_3$, implying that $\mu_1 > \mu_2$. Actually, all of these computations are unnecessary. Since the comparable confidence intervals of $\underline{CI}_1$ = (23.18, 27.82) and $\underline{CI}_2$ = (17.22, 22.78) are separable, and since $\underline{M}_1 > \underline{M}_2$, one can decide immediately that μ_1 is significantly larger than μ_2 at $\alpha/2$ and this conclusion is made with a Type III error probability less than 2.5%. Suppose the effect size $\delta_A = 3$ and since $\underline{SE}$ = 10 {(1/36) + (1/25)} = 2.603, we have $\underline{Z}_2$ = 1.96 - (3/2.603) = .807, $\underline{Z}_1$ = - (1.96 + (3/2.603)) = - 3.112 so that $\Phi(\underline{Z}_1) = \Phi(-3.112) = .000927$, $\Phi(\underline{Z}_2) = \Phi(.807) = .7903$ and

$$\pi(\delta) = 1 - .7903 + .000927 = 0.21063,$$

and

$$\pi(\delta)^* = \pi(\delta) - \gamma = 0.21063 - .00927 = .2097.$$

Hence, the estimate of Type III error rate is about .93%.

Comments

In this procedure, both the two within-sample tests ($\underline{H}_{0,i}$, $\underline{i}$ = 1, 2) and the one-tailed tests for mean discrepancy ($\underline{H}_{d,i}$, $\underline{i}$ = 1, 2) can have asymmetric regions as long as $\alpha_1 + \alpha_2 = \alpha$ (for $\underline{H}_{0,i}$, $\underline{i}$ = 1, 2) and $\alpha_{d1} + \alpha_{d2} = \Gamma$ (for $\underline{H}_{d,i}$, $\underline{i}$ = 1, 2), where Γ is a function of $\alpha_1 + \alpha_2$ as shown previously. The same statements of the decision rules will apply with the appropriate

changes in the significance levels. In the symmetrical simultaneous tests, the maximum probability of Type III error is $\alpha/2$. On the other hand, in the asymmetric simultaneous tests, it is equal to $\Gamma/2$, where $\alpha/2 \leq \Gamma/2 < \Gamma \leq \alpha$ (Shaffer, 1972). In the extreme asymmetric case, if one of the α_i in the test of $\underline{H}_{0,i}$ is set to zero then the other is set to equal α. At the same time, if one of the α_{di} is set at zero then Γ is also equal to α. In other words, the within-sample and between-sample tests are reduced to two independent one-tailed tests, each is evaluated at α). Hence, the symmetric test has an advantage that it minimizes the maximum probability of a Type III error. However, it is not necessary and not always best to impose symmetric critical regions since it may be more important to detect the effect of one sample, or differences in one direction, than in the other (Kaiser, 1960, p. 166; Shaffer, 1972, p. 196).

For the three-choice hypothesis of $\underline{H}_1$: $\mu_1 - \mu_2 = 0$, $\underline{H}_2$: $\mu_1 > \mu_2 = 0$, $\underline{H}_3$: $\mu_1 < \mu_2$, Hand, McCarter, and Hand (1985) proposed a procedure for testing the directional two-tailed hypothesis by just evaluating the $100(1 - \alpha)\%$ confidence of $\mu_1 - \mu_2$. Their decision rules are: "(a) if the signs of the two limits are different (zero is in the interval), then refuse to reach any conclusion about the population difference, (b) if both signs are positive, then accept H_1 and reject both $\underline{H}_0$ and $\underline{H}_2$, and (c) if both signs are negative, then accept $\underline{H}_2$ and reject both $\underline{H}_0$ and $\underline{H}_1$" (p. 495). Certainly, this decision rule can replace our rule (ii) since they imply the same statistical outcomes. However, for a simultaneous evaluation of one-sample parameter estimation and two-sample statistical inference, the need for comparable confidence intervals and hence, our decision rule set of (i) and (ii) are well-grounded.

Using Comparable Confidence Intervals for Pre-planed Multiple Comparisons

General Framework

If there are few a priori contrasts to be tested, the simplest method is the use of multiple t tests (Howell, 1997, p. 354). The procedure for simultaneous parameter estimation and statistical inference developed above is applicable to testing linear contrasts of means, proportions, and correlations under both conditions of variance homogeneity and heterogeneity. However, there are three necessary modifications for this purpose. First, under the assumption of variance homogeneity in an experiment of k independent groups, the pooled variance (S'^2) in computing the standard errors of the estimates (SE_i, $i = 1, 2, ..., k$) is replaced by the error mean square (MSE) obtained from the one-way ANOVA table for the total sample. The formula for MSE is specified as

$$MSE = \sum_{j=1}^{k}(n_j - 1)S^2_j \Big/ \sum_{j=1}^{k}(n_j - 1), \tag{15}$$

(Marascuilo and Serlin, 1988, p. 433), where S^2_j is the jth sample variance. Secondly, summary statistics of the specified contrasts must be computed in terms of weighted values. For example, in testing H_0: $\mu_{A+B} = \mu_C$, the weighted mean and variance for the combined group of A and B are

$$M_{AB} = (n_A M_A + n_B M_B)/(N_{AB} - 2) \tag{16}$$

and,

(17) $$\underline{S}^2_{AB} = [(\underline{n}_A - 1)\underline{S}^2_A + (\underline{n}_B - 1)\underline{S}^2_B]/(\underline{N}_{AB} - 2)$$

where $\underline{N}_{AB} = \underline{n}_A + \underline{n}_B$; $\underline{M}_A$ and $\underline{M}_B$ represent the sample means, and $\underline{S}^2_A$ and $\underline{S}^2_B$, the sample variances, of the two groups A and B, respectively. Thirdly, since each contrast represents a hypothesis to be tested using the same total sample, Student's $\underline{t}$ statistics are replaced by Bonferroni-Dunn's $\underline{t}$ to control for the familywise error rate (Howell, 1997, p. 362-364; Marascuilo and Serlin, 1988, Chapter 33). Several authors (e.g., Dayton and Schafer, 1973, and Schafer, 1992) have recommended the universal use of the Bonferroni adjustment for controlling familywise error rate in multiple comparisons as well as tests of correlations and proportions because the procedure is simple and requires no restrictions on the nature of the dependence of the tests. Moreover, comparing to other more difficult and restrictive methods, the loss of the statistical power due to the Bonferroni adjustment is minimal.

Preplanned Multiple Comparison of Means

Suppose the population values for all group means and variances are unknown. A researcher would like to test simultaneously $\underline{h}$ hypotheses, ($\underline{h} < \underline{k}$), on the contrasts $\Sigma\underline{a}_{j,d}\mu_j = 0$ ($\underline{d} = 1, 2, ..., \underline{h}$), where $\underline{a}_j = 1, -1$ or 0 ($\underline{j} = 1, 2, ..., \underline{k}$) = the $\underline{j}$th contrast weight such that $\Sigma\underline{a}_j = 0$, and Σ = the sum over $\underline{k}$ groups. Suppose the researcher is also interested in knowing the range of μ_j ($\underline{j} = 1, ..., \underline{k}$). In other words, the researcher wishes to test the following hypotheses:

($\underline{k}$ within-sample tests): $\underline{H}_{0,j}$: $\mu_j = \mu_0$ vs. $\underline{H}_{A,j}$: $\mu_j \neq \mu_0$, $\underline{j} = 1, 2, ..., \underline{k}$; where μ_0 can be any value of interest.

($\underline{h}$ across-sample tests): $\underline{H}_{0,d}$: $\Sigma\underline{a}_{j,d}\mu_{j,d} = 0$ versus $\underline{H}_{A,d}$: $\Sigma\underline{a}_{j,d}\mu_{j,d} \neq 0$, $\underline{d} = 1, 2, ..., \underline{h}$, where $\Sigma\underline{a}_{j,d}\mu_{j,d}$ is the $\underline{d}$th contrast involving $\underline{m}$ group means with nonzero contrast coefficients.

Each of the k within-sample tests is conducted the same way as discussed above for the one-sample tests of means. Given that the groups are independent, there is no familywise error rate to be protected. On the other hand, the procedure for the h across-sample tests is not so simple. Let us first consider the case in which the assumption of variance homogeneity is tenable across all k groups. The 100(1 - α)% two-tailed confidence interval for the dth contrast is expressed as

(18) $$CI_d: |\Sigma a_{j,d} M_{j,d}| \pm t_{f,\, 1 - \alpha/2h} SE_d,$$

for $d = 1, 2, ..., h$, where $|x|$ = the absolute value of x, $SE_d = \sqrt{\{MSE\Sigma(a_{j,d}^2/n_j)\}}$, $N = n_1 + ... + n_k$ = the total sample size, $f = N - k$ = the degree of freedom of the t statistic, $\alpha/2h$ = the familywise Type I error rate according to the Dunn-Bonferroni adjustment for h simultaneous hypotheses, and MSE is given in Equation 15.

For each contrast in the h across-sample tests, there are m within-sample tests, ($m \leq k$), corresponding to the m nonzero contrast coefficients among a_j, $j = 1, 2, ..., k$. Suppose μ_i ($i = 1, ..., m$) is included in the specification of the dth contrast. The correction factor for computing the comparable confidence interval for μ_i is of the form

$$c_{f,d} = SE_d/SE_c = \sqrt{[MSE\Sigma(a_{j,d}^2/n_j)]}/\Sigma\sqrt{\{S_j^2/n_j\}},$$

where $SE_d = \sqrt{\{MSE(a_{1,d}^2/n_1 + ... + a_{k,d}^2/n_k)\}}$, ($k$ - m) coefficients of a_j are zero, and $SE_c = S_1/\sqrt{n_1} + ... + S_k/\sqrt{n_k}$. Then the 100(1 - α)% two-tailed comparable confidence interval for $\mu_{i,d}$, $i = 1, ..., m$, is given by

(19) $$CI^*_{i,d}: M_i \pm t_{f,\, 1 - \alpha/2h} SE^*_i,$$

where $\underline{SE}^*_i = \underline{c}_{f,d}\underline{SE}_i$. The associated comparative significance level is computed as

$$\alpha' = 2[1 - \underline{T}_f(\underline{c}_f\underline{t}_{f,\,1\,-\,\alpha/2h})].$$

where $\underline{T}_f(x) = Pr(\underline{X} \leq \underline{x})$ = the probability of obtaining a value of $\underline{x}$ for the $\underline{t}_{f,\,1\,-\,\alpha/2h}$ distribution.

In applying the above procedure under variance heterogeneity, one need to modify the degree of freedom for the $\underline{t}$ statistic. It is the solution for the equation $\underline{f}^* = \Sigma(\underline{n}_j - 1)[1 - (\underline{W}_j/\underline{W})]$, where $\underline{W}_j = \underline{n}_j/\underline{S}^2_j$ and $\underline{W} = \Sigma\underline{W}_j$ (according to the Welch-Aspin approximation. See Marascuilo and Serlin, 1988, pp. 435-437). In evaluating the comparable confidence intervals per contrast under both variance conditions, directional two-tailed hypotheses can be assessed just as in the case of simple tests of means difference.

Example 5

In a certain study on the influence of professional training on attitudes toward persons with disabilities, 71 subjects were randomly assigned into six experimental conditions ($\underline{n}_A$ = 11, $\underline{n}_B$ = 12, $\underline{n}_C$ = 10, $\underline{n}_D$ = 14, $\underline{n}_E$ = 11, $\underline{n}_F$ = 13), which were subsequently combined into four treatment groups; with $\underline{n}_1 = \underline{n}_A + \underline{n}_B$, $\underline{n}_2 = \underline{n}_C + \underline{n}_D$, $\underline{n}_3 = \underline{n}_E$ and $\underline{n}_4 = \underline{n}_F$. Responses on the variable "Bias themes"($\underline{X}$) were measured (e.g., as defined in Kemp and Mallinckrodt, 1996). Before data collection, the researcher was interested in testing four contrasts: μ_1 vs. μ_2, μ_1 vs. μ_3, μ_1 vs. μ_4, and μ_3 vs. μ_4. The mean square error ($\underline{MSE}$) for $\underline{X}$ computed for the four groups is found to be 0.459. The assumption of variance homogeneity is tenable for the four contrasts ($\underline{p} > .26$, $\underline{p} > .47$, $\underline{p} > .19$ and $\underline{p} > .25$, respectively) according to a test of variance homogeneity (Hinkle, Wiersma & Jurs, 1994, pp. 242-244; Ferguson & Takane, 1989, pp. 202-204). The results of the procedure for testing these contrasts are summarized in Table 3.

Insert Table 3 about here

In Panel 1 of Table 3, the sample means and standard deviations are computed according to Equations 16 and 17, respectively. The conventional $\underline{t}$ tests, with degrees of freedom $\underline{f}_j = \underline{n}_j - 1$ ($j = 1, 2, 3, 4$) and the $100(1 - \alpha)\%$ two-tailed conventional confidence intervals are computed with the critical values of $\underline{t}_{fj,\ 1\ -\ \alpha/2}$, where $\alpha = .05$. All of these individual confidence intervals overlap to each other.

In Panel 2, the $\underline{t}$ tests of means difference are conducted with degrees of freedom $\underline{f}_d = \underline{N} - \underline{k} = 71 - 4 = 67$, and evaluated at $\alpha^* = \alpha/\underline{h} = .05/4 = .0125$ (two-tailed). The critical value of the test statistic according to the Bonferroni-Dunn approximation is $\underline{t}_{71,\ .05/4} = 2.567$ (two-tailed) (obtained by extrapolation from Table 1, Appendix $\underline{t}$', Howell, 1997, p. 687; or by the command TINV(.0125/2, 67) in the SAS computer program). For a numerical illustration, let us consider the evaluation of the third contrast, G1 vs. G4. The contrast coefficients are specified as $\underline{a}_1 = 1$, $\underline{a}_2 = 0$, $\underline{a}_3 = 0$, and $\underline{a}_4 = -1$. Hence, upon diving the contrast mean ($\Sigma \underline{a}_{j,3}\underline{M}_{j,d3} = 1.54 - 1.74 = - .20$) by its standard error ($\underline{SE}_{d3} = \sqrt{\{\underline{MSE}\Sigma \underline{a}_{j,d3}^2/\underline{n}_j\}} = \sqrt{\{.0438[(1)^2/23 + (-1)^2/13]\}} = .0726$, one obtains the test statistic $\underline{t}_{d3} = - .20/.0726 = - 2.711$. The corresponding 95% confidence interval according to Dunn-Bonferoni approximation for the contrast is equal to

$$\underline{CI}_{d3}: \mid \Sigma \underline{a}_{j,d3}\underline{M}_{j,d3} \mid \pm \underline{t}_{f,\ 1\ -\ \alpha/2h}\underline{SE}_{d3} = \mid -2.71 \mid \pm (2.567)(.0726) = (-.383, -.010),$$

implying that the third contrast is statistically significant at $\alpha = .05$ (two-tailed). Since $\underline{SE}_c = \Sigma\sqrt{\{\underline{S}^2_j/\underline{n}_j\}} = \sqrt{\{(0.23)^2/23 + (0.20)^2/24 + (0.22)^2/11 + (0.18)^2/13\}} = .104$, the correction factor is given by

$$\underline{c}_{f,d3} = \underline{SE}_{d3}/\underline{SE}_c = .0726/.104 = .698$$

For the individual standard errors of $\underline{SE}_1 = 0.23/\sqrt{23} = .048$ and $\underline{SE}_4 = 0.18/\sqrt{13} = .050$, their corrected values are $\underline{SE}^*_{1,d3} = \underline{c}_{f,d3}\underline{SE}_1 = (.698)(.048) = .033$ and $\underline{SE}^*_{4,d3} = \underline{c}_{f,d3}\underline{SE}_4 = (.698)(.050) = .035$. Hence, in Panel 3, the 95% comparable confidence intervals for μ_1 and μ_4 in the third contrast are

$$\underline{CI}^*_{1,d3}: \underline{M}_1 \pm \underline{t}_{f=71,\ 1\ -\ \alpha/2h}\underline{SE}^*_{1,d3} = 1.54 \pm (2.567)(.033) = (1.46,\ 1.63),$$

and

$$\underline{CI}^*_{4,d3}: \underline{M}_4 \pm \underline{t}_{f=71,\ 1\ -\ \alpha/2h}\underline{SE}^*_{4,d3} = 1.74 \pm (2.567)(.035) = (1.65,\ 1.83),$$

respectively. As expected, the two comparable confidence intervals are separable. Moreover, at the risk of Type III error less than $\alpha^*/2 = .0125$, one can assume that $\mu_1 < \mu_4$. The corresponding comparative Type I error rate is

$$\alpha' = 2[1 - \underline{T}_{f=67}\{.698(2.567)\}] = 2[1 - .961] = .078,$$

implying that the 92.2% conventional confidence intervals for the means μ_1 and μ_4 are separable.

Comments

The above method is applicable to a-priori multiple comparisons of contrasts with different weights and symmetric critical regions. Extensions to the asymmetric critical regions are being studied. However, tables for Dunn-Bonferroni adjustment for t tests with unequal allocation of Type I error rates are available (Dayton and Schaffer, 1973).

Discussions and Conclusions

The paper has begun with a discussion on the logic and motivation for undertaking a study of comparable confidence intervals for two-sample means tests. The proposed procedures were then generalized in terms of parameter type, variance conditions and tailed error allocation for critical regions. Procedures for nondirectional and directional two-tailed tests for a simultaneous evaluation of within-sample parameter estimation and across-sample tests of parameter discrepancy for pre-planned simple and multiple comparisons were studied. It was suggested that for such a simultaneous evaluation, one need only to compute and analyze the comparable confidence intervals per parameter pair. Thus if the comparable confidence intervals are separable, one can assume that the pair of parameter estimates are statistically significant, proceeds to determine the confidence limits for such a difference to happen and makes prediction on the direction of the parameter difference with the risk of getting the direction wrong less than half the nominal significance level (with Dunn-Bonferroni adjustment in the case of multiple comparisons).

Researchers sometimes attempt to cover all bases by conducting a significance test and, when results are significant, calculating a confidence interval to estimate the parameter range. Unfortunately, making the decision to estimate parameter size contingent on the outcome of the significance test produces a biased estimate of the parameter (Schmidt, 1992) and loss of statistical power (Bancroft, 1944, Bennett, 1952). It is important to emphasize that the proposed procedures are developed for a-priori pairwise or multiple comparisons. Therefore, neither the construction nor the statistical power of comparable confidence intervals are influenced by the knowledge of the significance test outcomes. Because of these

concerns, comparable confidence intervals are not recommended for post-hoc pairwise comparisons.

A question of interest is how the power of comparable confidence intervals can be determined. The within-sample tests for ϕ_i, and the across-sample tests for $\phi_i \neq \phi_{i'}$, $i \neq i' = 1, 2, ..., k$, ($k \geq 2$), may have different statistical powers. Depending on the disparity in sample sizes and variances as well as the magnitude of ϕ_0, the within-sample tests could be less (or more) powerful than the across-sample tests for the same parameter pair. For example, in two-sample studies, if both ϕ_1 and ϕ_2 are small but substantially different and ϕ_0 is set at zero, it is likely that the between-sample test will lead to the rejection of the null hypothesis when it is false more often than the between-sample tests will at the same α level. However, the difference in power of within-sample and between-sample tests does not influence the separability of comparable confidence intervals. Conceptually, as far as the decisions on parameter discrepancy (or separability) and predicted direction of the disparity, the powers of comparable confidence intervals are akin to the powers of the across-sample tests.

As another point of clarification, in the procedures of a single pairwise comparison for two-sample studies, the pre-determined α level in the simultaneous evaluation of the within-sample and between-sample tests by two comparable confidence intervals does not require any adjustment of the type required for familywise error rate. Under each of the within-sample and between-sample tests in the proposed procedures, the chosen α represents the error rate per comparison (Howell, 1997, p. 349). When the two comparable confidence intervals are evaluated, α becomes the error rate per experiment (Howell, 1997, p. 349) for an experiment in which only one test has been conducted, namely the between-sample test. Of course, the adjustment for controlling familywise error rate is needed in a-priori multiple

comparisons.

For most practical research settings, it is believed that the proposed methods are both effective and meaningful. Instead of conducting three significance tests ($\underline{H}_{0i}$, $\underline{i}$ = 1, 2 and $\underline{H}_d$), only two comparable confidence intervals are required. The saving can be more substantial with respect to a simultaneous testing of individual hypotheses and multiple comparisons. With regards to the interpretation issue, one may question about the appropriateness of using interval separability to make decisions on statistical significance. We recognize that the separability of two comparable confidence intervals and the statistical significance of a test for parameter discrepancy may have different meanings to researchers. However, confidence intervals posses the same mathematical properties of significance tests, and more. A confidence coefficient represents the probability of producing an interval containing the true value of the parameter of interest. One declaring that the two confidence intervals are disjointed, with specified confidence coefficient and interval bounds, when the influence of variations in sample size and sample variances have been accounted for, conveys a clearer message about the magnitude and nature of difference for the two sample estimates than saying that their difference is "significant". This is because the probability that the test statistic would take a value as extreme or more extreme than actually observed is smaller than $\alpha/2$ does not depict the whole picture of the difference. The confidence widths actually estimate the relative sizes of the individual effects. The measure of separability (or overlapping) of the two comparable confidence intervals indicates the size of their discrepancy (or similarity) or an estimate of the effect size of the difference scores.

Comparing to the conventional confidence intervals, the proposed method yields confidence intervals with narrower widths, overlapping bounds for insignificant means

differences and separable limits for statistically significant results. These properties are confirmed in the examples under consideration. The proposed procedure is particularly useful whenever it is meaningful to evaluate confidence intervals for both parameter estimation and hypothesis testing. For example, in the calibration of IQ scores, psychometricians may want to determine the score range of IQ groups such that the means difference between any pair of adjacent groups will be statistically significant. Psychiatrists are also interested in testing the difference of Verbal IQ score (VIQ) against Performance IQ score (PIQ) as well as the individual ranges of VIQ and PIQ that the potential patients may belong to. As another area of potential application, studies for decision-making purposes generally require both statistical estimation and statistical inference, as opposed to exploratory studies which are mainly based on hypothesis testing for explanation purposes. For example, in clinical research, explanatory trials are conducted to determine whether a difference in treatments exists at all whereas the more sophisticated management trials is aimed at not only comparing treatment means but also deciding which treatment are better or should be used (Willan, 1994).

Cox and Hinkley (1974) considered interval estimation as "the central problem of statistical inference" and Cox (1977) concluded "that therefore estimation, at least roughly, of the magnitude of effects is in general essential regardless of whether statistically significant departure from the null hypothesis is achieved" (p.70). Tukey (1991) discussed four compelling reasons about the importance of confidence intervals and maintained, "It should be clear (a) that confidence intervals are irreplaceable and (b) why they are needed" (p. 102). The proposed methods serves to illustrate, and advance, the utility of the confidence interval approach in supporting these arguments.

Footnotes

[1] Researchers believing that the magnitude of effect size to be irrelevant as long as the effect is either statistically or practically important may not be interested in the comparison of confidence interval widths (Parker, 1995; Prentice and Miller, 1992). However, our concern is on the waste of one-sample information if one only conducts the evaluation of the difference scores.

[2] The preference for confidence intervals has been echoed through diverse disciplines (e.g., in the behavioral sciences by Carver, 1978, 1993; Cohen, 1994; LaForge, 1967; Schmidt, 1996; Serlin, 1993 and Shaffer, 1995; engineering by Hahn, 1974; Hahn and Meeker, 1991; Hsu, 1996; and Natrella, 1960; and medical studies by Borenstein, 1994; Gardner and Altman, 1986; Langman, 1986; Poole, 1987; Rothman, 1978, 1986; and Thompson, 1987. In fact, this list is severely incomplete). See also the comments on Cohen (1994) by Baril and Cannon (1995), Cohen (1995), Frick (1995), Hubbard (1995), McGraw (1995), Parker (1995), and Svyantek and Ekeberg (1995). For a critical view on the role of confidence intervals in hypothesis testing, see Cortina and Dunlap (1997).

[3] A full treatment of using confidence intervals for multiple comparisons is found in Hsu (1996).

[4] For the test of $\underline{H}_{0d}$, the test statistic is equal to 2.1126, corresponding to a p-value of $\alpha_p = 2*(1 - \Phi(2.1126)) = .0346$ and the corresponding significant confidence interval is $\underline{CI}_s = (0, 11)$.

[5] If the test statistic for a conventional two-tailed hypothesis is not significant at $\alpha/2$ then the Type III error for the corresponding two-tailed directional test is zero or undefined,

and if the test statistic is significant at $\alpha/2$, as the parameter discrepancy (δ), and power of the test, decreases to zero, the Type III error increases to its maximum value of $\alpha/2$. On the contrary, the Type III error becomes infinitesimal as power increases to 1

References

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. Annals of Mathematical Statistics, 15, 190-204.

Baril, G. L. & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless. American Psychologists, 50, 1098-1099.

Borenstein, M. (1994). A note on the use of confidence intervals in psychiatric research. Psychopharmacology Bulletin, 30, 236-238.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologists, 49, 997-1003.

Cohen, J. (1995). The earth is round ($p < .05$). Rejoinder. American Psychologists, 49, 1103.

Cortina, J. M. & Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Cox, D. R. (1977). The role of significance tests. Scandinavian Journal of Statistics, 4, 49-70.

Cox, D. R. & Hinkley, D. V. (1974). Theoretical Statistics. London: Chapman & Hall.

Bennett, B. M. (1952). Estimation of means on the basis of preliminary tests of significance. Annals of the Institute of Mathematical Statistics, 4, 31-43.

Dayton, C. M. & Schafer, W. D. (1973). Extended tables of t and chi square for Bonferroni tests with unequal error allocation. Journal of the American Statistical Association, 68, 78-83.

DeGroot, M. H. (1975). Probability and Statistics. Menlo Park, CA: Addison Wesley.

Ferguson, G. A. & Takane, Y. (1989). Statistical analysis in Psychology (6th ed.). New York, NY: McGraw-Hill.

Frick, R. W. (1995). A problem with confidence intervals. American Psychologists, 50, 1102-1103.

Gardner, M. J. & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal, 292, 746-750.

Hand, J., McCarter, R. E. & Hand, M. R. (1985). The procedures and justification of a two-tailed directional test of significance. Psychological Reports, 56, 495-498.

Hays, W. L. (1988). Statistics (4th ed.). New York: Holt, Rinehart & Winston.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston, MA: Houghton Mifflin.

Howell, D. C. (1997). Statistical methods for Psychology. Belmont, CA:Duxbury Press.

Hsu, J. C. (1996). Multiple comparisons. Theory and methods. London: Chapman & Hall.

Hubbard, R. (1995). The earth is highly significantly round ($p < .0001$). American Psychologists, 50, 1098.

Kaiser, H. F. (1960). Directional statistical decisions. Psychological Review, 67, 160-167.

Kemp, N. T. & Mallinckrodt, B. (1996). Impact of professional training on case conceptualization of clients with a disability. Professional Psychology Research and Practice, 27, 378-385.

LaForge, R. (1967). Confidence intervals or tests of significance in scientific research? Psychological Bulletin, 68, 446-447.

Langman, M. J. S. (1986). Towards estimation and confidence intervals. British Medical Journal, 292, 716.

Leventhal, L. & Huynh, C-L. (1996a). Directional decisions for two-tailed tests: Power, error rates, and sample size. Psychological Methods, 1, 278-292.

Leventhal, L. & Huynh, C-L. (1996b). Analyzing listening tests with directional two-tailed tests. Journal of the Audio Engineering Society, 44, 850-863.

McGraw, K. O. (1995). Determine false alarm rates in null hypothesis testing research. American Psychologists, 50, 1099-1100.

Natrella, M. G. (1960). The relation between confidence intervals and tests of significance. American Statistician, 14, 20-22.

Parker. S. (1995). The "difference of means" may not be the "effect size". American Psychologists, 50, 1101-1102.

Poole, C. (1987). Beyond the confidence interval. American Journal of Public Health, 77, 195-199.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 112, 160-164.

Rothman, K. J. (1978). A show of confidence. New England Journal of Medicine, 299, 1362-1363.

Rothman, K. J. (1986). Significance questing. Annals of Internal Medicine, 3, 445-447.

SAS Institute Inc. (1990). SAS language, Version 6 (4th ed.). Cary, NC: Authors.

Satterwaite, F. W. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 110-114.

Schafer, W. D. (1992). Simultaneous inference options for statistical decision making (Editorial). Measurement and Evaluation in Counseling and Development, 25, 98-101.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47, 1173-1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 15-129.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61, 350-360.

Shaffer, J. P. (1972). Directional statistical hypotheses and comparisons among means. Psychological Bulletin, 77, 195-197.

Shaffer, J. P. (1972). Multiple hypothesis testing. Annual Review of Psychology, 46, 561-584.

Svyantek, D. J. & Ekeberg, S. E. (1995). The earth is round (So we can probably get there from here). American Psychologist, 50, 1101.

Thompson, W. D. (1987). Statistical criteria in the interpretation of epidemiologic data. American Journal of Public Health, 77, 191-194.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.

Welch, B. L. (1938). The significance of the difference between two means when the

population variances are unequal. Biometrika, 29, 350-362.

Welsh, A. H. (1996). Aspects of statistical inference.

Willan, A. R. (1994). Power function arguments in support of an alternative approach for analyzing management trials. Controlled Clinical Trials, 15, 211-219.

Zehna, P. W. (1970). Probability Distributions and Statistics. Boston, MA: Allyn & Bacon.

Table 1

Correction Factors in Computing Confidence Intervals for Two-sample Tests of Means (μ), Proportions (π) and Correlation Coefficients (ρ)

Case	Null Hypothesis	Sampling Distribution[1]	Distribution Conditions	Correction Factor
1a	H_0: $\mu_1 - \mu_2 = 0$	$\underline{Z}$	$\sigma^2_1 = \sigma^2_2$	$\underline{c}_f = \frac{\sqrt{\{1 + \underline{m}\}}}{1 + \sqrt{\underline{m}}}$
1b	H_0: $\mu_1 - \mu_2 = 0$	$\underline{Z}$	$\sigma^2_1 \neq \sigma^2_2$	$\underline{c}_f = \frac{\sqrt{\{\underline{u} + \underline{m}\}}}{\sqrt{\{\underline{u} + \sqrt{\underline{m}}}$
1c	H_0: $\mu_1 - \mu_2 = 0$	$\underline{t}(\underline{f})$	$\sigma^2_1 = \sigma^2_2$	$\underline{c}_f = \frac{\sqrt{\{(1 + \underline{m})(1 + \underline{g}\underline{v})\}}}{\sqrt{\{1 + \underline{g}\}(\sqrt{\underline{m}} + \sqrt{\underline{v}})}}$
1d	H_0: $\mu_1 - \mu_2 = 0$	$\underline{t}(\underline{f}')$	$\sigma^2_1 \neq \sigma^2_2$	$\underline{c}_f = \frac{\sqrt{\{\underline{v} + \underline{m}\}}}{\sqrt{\underline{v}} + \sqrt{\underline{m}}}$
2	H_0: $\pi_1 - \pi_2 = 0$	$\underline{Z}$	$\underline{n}_i p_i > 5$ $\underline{n}_i(1 - p_i) > 5$	$\underline{c}_f = \frac{\sqrt{\{\underline{S}^{2'} + \underline{m}\underline{S}^{2'}\}}}{\sqrt{\underline{S}^2_2 + \sqrt{\{\underline{m}\underline{S}^2_1\}}}}$
3	H_0: $\rho_1 - \rho_2 = 0$	$\underline{Z}$		$\underline{c}_f = \frac{\sqrt{\{(\underline{n}_1+3) + (\underline{n}_2+3)\}}}{\sqrt{\{\underline{n}_1+3\}} + \sqrt{\{\underline{n}_2 + 3\}}}$

Note. $\underline{m} = \underline{n}_2/\underline{n}_1$, $\underline{u} = \sigma^2_2/\sigma^2_1$, $\underline{v} = S^2_2/S^2_1$, $\underline{f} = \underline{f}_1 + \underline{f}_2$, $\underline{f}_i = \underline{n}_i - 1$ for i = 1, 2, $\underline{f}' = [(\underline{w}^2/\underline{f}_1) + ((1 - \underline{w})^2/\underline{f}_1)]^{-1}$, $\underline{w} = \underline{S}^2_1/\underline{n}_1\underline{q}$ and $\underline{q} = (S^2_1/\underline{n}_1) + (S^2_2/\underline{n}_2)$, $\underline{S}^{2'} = \{\underline{p}'(1 - \underline{p}')\}$, $\underline{p}' = (\underline{n}_1\underline{p}_1 + \underline{n}_2\underline{p}_2)/\underline{N}$, and $\underline{N} = \underline{n}_1 + \underline{n}_2$.

[1] Sampling distribution of the test statistics.

Table 2

Confidence Coefficients of Conventional and Comparable Confidence Intervals Corresponding to Selected Values of Type I Error and Correction Factor

Panel 1. Z tests

			Correction factor (c_f)										
α	CI*	Z_p	0.70	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.80
.0001	.999	3.891	0.994	0.994	0.995	0.996	0.996	0.997	0.997	0.997	0.998	0.998	0.998
.01	.99	2.576	0.929	0.933	0.936	0.940	0.943	0.947	0.950	0.953	0.955	0.958	0.961
.05	.95	1.960	0.830	0.836	0.842	0.848	0.853	0.858	0.864	0.869	0.874	0.878	0.883
.10	.90	1.645	0.750	0.757	0.764	0.770	0.776	0.783	0.789	0.795	0.801	0.806	0.812
.20	.80	1.282	0.630	0.637	0.644	0.651	0.657	0.664	0.670	0.676	0.685	0.689	0.695

Panel 2. t Tests

				Correction factor (c_f)									
α	CI*	f	T_p	0.70	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.80
.01	.99	10	3.169	0.949	0.952	0.955	0.957	0.959	0.961	0.963	0.965	0.967	0.969	0.970
		20	2.845	0.940	0.943	0.946	0.949	0.952	0.955	0.957	0.960	0.962	0.964	0.966
		40	2.704	0.934	0.938	0.941	0.945	0.948	0.951	0.954	0.956	0.959	0.962	0.963
		80	2.639	0.932	0.935	0.939	0.942	0.946	0.949	0.952	0.955	0.957	0.960	0.962
		160	2.607	0.930	0.934	0.938	0.941	0.944	0.948	0.951	0.954	0.956	0.959	0.962
.05	.95	10	2.228	0.850	0.855	0.860	0.865	0.870	0.874	0.879	0.883	0.887	0.891	0.895
		20	2.086	0.840	0.846	0.851	0.857	0.862	0.867	0.871	0.876	0.881	0.885	0.889
		40	2.021	0.835	0.841	0.847	0.852	0.857	0.863	0.868	0.872	0.877	0.882	0.886
		80	1.991	0.833	0.838	0.844	0.850	0.855	0.861	0.866	0.871	0.875	0.880	0.885
		160	1.975	0.832	0.837	0.843	0.849	0.854	0.859	0.865	0.870	0.875	0.879	0.884
.10	.90	10	1.812	0.767	0.773	0.779	0.785	0.790	0.796	0.802	0.807	0.812	0.817	0.822
		20	1.725	0.759	0.765	0.771	0.777	0.784	0.789	0.795	0.801	0.806	0.812	0.817
		40	1.684	0.755	0.762	0.768	0.774	0.780	0.786	0.792	0.798	0.803	0.809	0.814
		80	1.664	0.752	0.759	0.766	0.772	0.778	0.785	0.790	0.796	0.802	0.808	0.813
		160	1.654	0.751	0.758	0.765	0.771	0.777	0.784	0.790	0.795	0.801	0.807	0.812
.20	.80	10	1.372	0.641	0.647	0.654	0.660	0.666	0.672	0.678	0.684	0.690	0.696	0.702
		20	1.325	0.635	0.642	0.649	0.655	0.662	0.668	0.674	0.680	0.686	0.692	0.698
		40	1.303	0.633	0.640	0.646	0.653	0.659	0.666	0.672	0.678	0.684	0.691	0.697
		80	1.292	0.632	0.638	0.645	0.652	0.658	0.665	0.671	0.677	0.683	0.690	0.696
		160	1.287	0.631	0.638	0.644	0.651	0.658	0.664	0.670	0.677	0.683	0.689	0.695

Note. α = Nominal Type I error, CI* = Confidence coefficient of the comparable confidence interval, Z_p = Critical value of the Z statistic for two-tail test at α level, t_p = Critical value of the t statistic of degree of freedom f for two-tail test at α level under the assumption of variance homogeneity, and f = n - 1 = degree of freedom for t tests.

Table 3

An Example of Planned Multiple Comparisons for Means (with hypothetical data)

Panel 1 Conventional One-sample Tests & Confidence Intervals					Panel 2 Test of Means Difference				Panel 3 Comparable Confidence Intervals Contrasts			
	n	M	t	CI_L	Contrast	d	t	CI_L		$\mu_1-\mu_2$ CI_L	$\mu_1-\mu_3$ CI_L	$\mu_1-\mu_4$ CI_L	$\mu_3-\mu_4$ CI_L
		SD	p<t	CI_U		SE	p<t	CI_U	Group	CI_U	CI_I	CI_U	CI_U
G1	23	1.54	32.34	1.44	$\mu_1 - \mu_2$	-.10	-1.65	-.257	G1	1.47	1.45	1.46	
		0.23	.000	1.64		.06	.104	.056		1.61	1.63	1.63	
G2	24	1.64	40.43	1.56	$\mu_1 - \mu_3$	0.01	0.17	-.184	G2	1.58			
		0.20	.000	1.73		.06	.866	.210		1.71			
G3	11	1.53	23.06	1.38	$\mu_1 - \mu_4$	.20	-2.71	-.383	G3		1.40		1.39
		0.22	.000	1.68		.07	.008	-.010			1.66		1.67
G4	13	1.74	34.85	1.63	$\mu_3 - \mu_4$	0.21	-2.45	-.430	G4			1.65	1.63
		0.18	.000	1.85		.09	.017	.010				1.83	1.84

Note. Values are rounded up to two or three decimals. M = Sample mean, SD = Sample standard deviation, d = Sample means difference, SE = Standard error of d, t = Bonferroni-Dunn's t test statistic, p < t = p-value of t for two-tailed tests, CI_L = Lower bound of the 95% confidence interval, CI_U = Upper bound of the 95% confidence interval.

Appendix

DERIVATION OF THE FORMULAS IN TABLE 1

For each procedure listed below, the formulas related to two-tailed tests under unbalanced designs ($n_i \neq n_j$, $i \neq j = 1, ..., k$) are given for the following substeps: (i) confidence interval for the difference scores (CI_d); (ii) one-sample comparable confidence intervals (CI^*); and (iii) comparative Type I error (α'). For one-tailed tests, replace $\alpha/2$ by α and α' by $\alpha'/2$ in the formulas. Based on theses results, formulas under balanced designs are trivial. Some of the notations that will be used repeatedly are: $m = n_2/n_1$, $u = \sigma^2_2/\sigma^2_1$, $v = S^2_2/S^2_1$, $f = f_1 + f_2$, $f_i = n_i - 1$ for $i = 1, 2$, $\Phi(x)$ and $T_{df}(x)$ represent the cumulative distribution functions (CDF) of standard normal, and t with degrees of freedom df, respectively, evaluated at x.

Case 1. Simple Comparisons of Means

1a. Variance homogeneity ($\sigma_1 = \sigma_2 = \sigma$, known)

(i) CI_d: $(M_1 - M_2) \pm Z_{1-\alpha/2}SE_d$, where $SE_d = \sigma\sqrt{\{(n_1 + n_2)/n_1n_2\}}$.

(ii) CI^*_i: $M_i \pm Z_{1-\alpha/2}SE^*_i$, where $SE^*_i = c_fSE_i$, $SE_i = \sigma/\sqrt{n_i}$, $i = 1, 2$, and $c_f = \sqrt{\{n_1 + n_2\}}/(\sqrt{n_1} + \sqrt{n_1}) = \sqrt{\{1 + m\}}[1 + \sqrt{m}]^{-1}$.

(iii) $\alpha' = 2[1 - \Phi(\sqrt{\{c_fZ_{1-\alpha/2}\}})]$.

1b. Variance heterogeneity ($\sigma_1 \neq \sigma_2$, known)

(i) CI_d: $(M_1 - M_2) \pm Z_{1-\alpha/2}SE_d$, where $SE_d = \sqrt{\{(\sigma^2_1/n_1) + (\sigma^2_1/n_2)\}}$.

(ii) $\underline{CI}^*_i$: $\underline{M}_i \pm \underline{Z}_{1-\alpha/2}\underline{SE}^*_i$, where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, $\underline{SE}_i = \sigma_i/\sqrt{\underline{n}_i}$, $\underline{i} = 1, 2,$

and $\underline{c}_f = \sqrt{\{\underline{n}_2\sigma^2_1 + \underline{n}_1\sigma^2_2\}[\sigma_1\sqrt{\underline{n}_2} + \sigma_2\sqrt{\underline{n}_1}]^{-1}} = \sqrt{\{\underline{m} + \underline{u}\}[\sqrt{\underline{m}} + \sqrt{\underline{u}}]^{-1}}$.

(iii) $\alpha' = 2[1 - \Phi(\sqrt{\{\underline{c}_f\underline{Z}_{1-\alpha/2}\}})]$.

1c. Variance homogeneity ($\sigma_1 = \sigma_2 = \sigma$, unknown)

(i) $\underline{CI}_d$: $(\underline{M}_1 - \underline{M}_2) \pm \underline{t}_{f,1-\alpha/2}\underline{SE}_d$, where $\underline{SE}_d = \sigma\sqrt{\{2/\underline{n}\}}$, $\underline{f} = \underline{f}_1 + \underline{f}_2$ and $\underline{f}_i = \underline{n}_i - 1$,

$\underline{i} = 1, 2.$

(ii) $\underline{CI}^*_i$: $\underline{M}_i \pm \underline{t}_{f,1-\alpha/2}\underline{SE}^*_i$, where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, $\underline{SE}_i = \underline{S}_i/\sqrt{\underline{n}_i}$, $\underline{i} = 1, 2,$ and

$\underline{c}_f = \sqrt{\{(1 + \underline{g}\underline{u})(1 + \underline{m})[1 + \underline{g}]^{-1}\}}/[\sqrt{\underline{m}} + \sqrt{\underline{u}}] = \sqrt{\{(1 + \underline{m})(1 + \underline{g}\underline{v})\}}[\sqrt{\{1 + \underline{g}\}}(\sqrt{\underline{m}} + \sqrt{\underline{v}})]^{-1}$

where $\underline{g} = \underline{f}_2/\underline{f}_1$.

(iii) $\alpha' = 2[1 - \underline{T}_f(\sqrt{\{\underline{c}_f\underline{t}_{f,1-\alpha/2}\}})]$.

1d. Variance heterogeneity ($\sigma_1 \neq \sigma_2$, unknown)

(i) $\underline{CI}_d$: $(\underline{M}_1 - \underline{M}_2) \pm \underline{t}_{f^*,1-\alpha/2}\underline{SE}_d$, where $\underline{SE}_d = \sqrt{\{(\underline{S}_1^2/\underline{n}_1) + (\underline{S}_2^2/\underline{n}_2)\}}$, and

$$\underline{f}^* = \frac{(\underline{S}_{M1}^2 + \underline{S}_{M2}^2)^2}{\dfrac{(\underline{S}_{M1}^2)^2}{\underline{n}_1 - 1} + \dfrac{(\underline{S}_{M2}^2)^2}{\underline{n}_2 - 1}} = [\underline{w}^2/\underline{f}_1 + (1 - \underline{w}^2)/\underline{f}_2]^{-1},$$

(Satterwaite, 1946; Welch, 1938), where $\underline{S}_{Mi}^2 = \underline{S}_i^2/\underline{n}_i$, $i = 1, 2$, $\underline{w} = \underline{S}^2_1/\underline{n}_1\underline{q}$ and $\underline{q} = (\underline{S}^2_1/\underline{n}_1) +$

$(\underline{S}^2_2/\underline{n}_2)$.

(ii) $\underline{CI}^*_i$: $\underline{M}_i \pm \underline{t}_{f^*,1-\alpha/2}\underline{SE}^*_i$, where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, $\underline{c}_f = [\sqrt{\underline{w}} + \sqrt{\{1 - \underline{w}\}}]^{-1}$. Since $\underline{w} = \underline{m}/(\underline{m} + \underline{v})$, the correction factor can be written as $\underline{c}_f = \sqrt{\{\underline{m} + \underline{v}\}}[\sqrt{\underline{m}} + \sqrt{\underline{v}}]^{-1}$.

(iii) $\alpha^* = 2[1 - \underline{T}_{f^*}(\sqrt{\{\underline{c}_f\underline{t}_{f^*,1-\alpha/2}\}})]$.

Case 2. Simple Comparisons of Proportions

In testing the null hypothesis: $H_{0,i}$: $\pi_i = \pi_0$, the application of the following procedure requires that $\underline{p}_i\underline{n}_i > 5$ and $\underline{n}_i(1 - \underline{p}_i) > 5$ ($\underline{i} = 1, 2$), where π_i and $\underline{p}_i$ are the population and sample proportions of the ith population, respectively.

(i) $\underline{CI}_i$: $\underline{p}_i \pm \underline{Z}_{1-\alpha/2}\underline{SE}_i$, where $\underline{SE}_i = S_i/\sqrt{\underline{n}_i}$ and $S_i = \sqrt{\{\underline{p}_i(1 - \underline{p}_i)\}}$, i = 1, 2.

(ii) $\underline{CI}_d$: $(\underline{p}_1 - \underline{p}_2) \pm \underline{Z}_{1-\alpha/2}\underline{SE}_d$, where $\underline{SE}_d = \underline{S}'\sqrt{\{\underline{N}/\underline{n}_1\underline{n}_2\}}$, $\underline{S}' = \sqrt{\{\underline{p}'(1 - \underline{p}')\}}$, $\underline{p}' = (\underline{n}_1\underline{p}_1 + \underline{n}_2\underline{p}_2)/\underline{N}$, and $\underline{N} = \underline{n}_1 + \underline{n}_2$.

(iii) $\underline{CI}^*_i$: $\underline{p}_i \pm \underline{Z}_{1-\alpha/2}\underline{SE}^*_i$, where $\underline{SE}^* = \underline{c}_f\underline{SE}_i$, $\underline{i} = 1, 2$, $\underline{SE}_i = S_i/\sqrt{\underline{n}_i}$, $S_i = \sqrt{\{\underline{p}_i(1 - \underline{p}_i)\}}$, and $\underline{c}_f = \underline{S}'\sqrt{\underline{N}[S_1\sqrt{\underline{n}_2} + \underline{S}_2\sqrt{\underline{n}_1}]^{-1}} = \sqrt{\{\underline{S}^{2'} + \underline{m}\underline{S}^{2'}\}}[\sqrt{\underline{S}^2_2} + \sqrt{\{\underline{m}\underline{S}^2_1\}}]^{-1}$.

(iv) $\alpha^* = 2[1 - \Phi(\sqrt{\underline{c}_f}\underline{Z}_{1-\alpha/2})]$.

Case 3. Simple Comparisons of Correlations

For statistical inference regarding ρ and ρ_i, where ρ denotes the population correlation coefficient, the following procedure requires the computation of $\underline{Z}_{ri}$, the Fisher's $\underline{Z}$ transformation of $\underline{r}_i$, i = 1, 2, where $\underline{r}_i$ represents the sample Pearson correlation of the ith population.

(a1) $\underline{CI}_i$ for $\underline{Z}_{\rho i}$: $\underline{Z}_{ri} \pm \underline{Z}_{1-\alpha/2}\underline{SE}_i = (\underline{Z}_{ri,L}, \underline{Z}_{ri,U})$, where $\underline{SE}_i = [\sqrt{\{n_i + 3\}}]^{-1}$, i = 1, 2; $\underline{Z}_{ri,L}$

and $\underline{Z}_{ri,U}$ are the lower and upper bounds of this confidence interval, respectively. They can be converted into the raw score units of measurement as $(\underline{r}_{i,L}, \underline{r}_{i,U})$: $([1 - \exp\{-2\underline{Z}_{ri,L}\}][1 + \exp\{-2\underline{Z}_{ri,L}\}]^{-1}, [1 - \exp\{-2\underline{Z}_{ri,U}\}][1 + \exp\{-2\underline{Z}_{ri,U}\}]^{-1})$.

(i) The $100(1 - \alpha)\%$ two-tailed confidence interval for $\underline{Z}_{\rho 1} - \underline{Z}_{\rho 2}$:

$$\underline{CI}_d: (\underline{Z}_{r1} - \underline{Z}_{r2}) \pm \underline{Z}_{1 - \alpha/2}\underline{SE}_d = (\underline{Z}_{rd,L}, \underline{Z}_{rd,U}),$$

where $\underline{SE}_d = \sqrt{\{(\underline{n}_1 + 3)^{-1} + (\underline{n}_2 + 3)^{-1}\}}$.

The $100(1 - \alpha)\%$ two-tailed confidence interval for $\rho_1 - \rho_2$:

$$\underline{CI}_d: ([1 - \exp\{-2\underline{Z}_{rd,L}\}][1 + \exp\{-2\underline{Z}_{rd,L}\}]^{-1}, [1 - \exp\{-2\underline{Z}_{rd,U}\}][1 + \exp\{-2\underline{Z}_{rd,U}\}]^{-1}).$$

(ii) The $100(1 - \alpha)\%$ two-tailed comparable confidence interval for $\underline{Z}_{\rho 1} - \underline{Z}_{\rho 2}$:

$$\underline{CI}^*_i: \underline{Z}_{ri} \pm \underline{Z}_{1 - \alpha/2}\underline{SE}^*_i = (\underline{Z}^*_{ri,L}, \underline{Z}^*_{ri,U}),$$

where $\underline{SE}^*_i = \underline{c}_f\underline{SE}_i$, $\underline{i} = 1, 2$, and

$$\underline{c}_f = \frac{\sqrt{\{\underline{n}1 + \underline{n}2 - 6\}}}{\sqrt{\{\underline{n}_1 - 3\}} + \sqrt{\{\underline{n}_2 - 3\}}}.$$

The $100(1 - \alpha)\%$ two-tailed comparable confidence interval for $\rho_1 - \rho_2$:

$$\underline{CI}^*_i: ([1 - \exp\{-2\underline{Z}^*_{ri,L}\}/[1 + \exp\{-2\underline{Z}^*_{ri,L}\}, [1 - \exp\{-2\underline{Z}^*_{ri,U}\}]/[1 + \exp\{-2\underline{Z}^*_{ri,U}\}])$$

(iii) $\alpha^* = 2[1 - \Phi(\sqrt{\underline{c}_f}\underline{Z}_{1 - \alpha/2})]$.

TM029111

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparable Confidence Intervals for Multi-sample and Replication studies

Author(s): CAM-LOI HUYNH, Ph. D.

Corporate Source: Paper presented at the 1998 Annual meeting of the American Educational Research Association, San Diego, CA

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ ☐	Level 2B ↑ ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries..

Sign here, →

Signature:	Printed Name/Position/Title: CAM-LOI HUYNH ASSOCIATE PROFESSOR
Organization/Address: DEPARTMENT OF PSYCHOLOGY UNIVERSITY OF MANITOBA, WINNIPEG, MB R3T 2N2, CANADA	Telephone: (204) 474 8260 FAX: (204) 474 7599 E-Mail Address: HUYNH@CC.UMANITOBA.CA Date: June 17, 1998

MICRO

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

KAREN SMITH
ACQUISITIONS COORDINATOR
ERIC/EECE
CHILDREN'S RESEARCH CENTER
51 GERTY DRIVE
CHAMPAIGN, ILLINOIS 61820-7469

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com